Alignments (pairwise)

pp1_alignments_1

Protein Prediction 1 - Protein structure
TUM summer 2014

© Burkhard Rost (TUM Munich)
Announcements

 Videos:  YouTube / www.rostlab.org

THANKS:

Tim Karl + Jonas Reeb

Special lectures:

• Apr 15 - Andrea Schafferhans

No lecture:

• Apr 17/22 Easter  
• May 01 Thu May day  
• May 06 Tue Student assembly  
• May 29 Thu Ascension day  
• Jun 10 Tue Whitsun holidays  
• Jun 19 Thu Corpus Christi

LAST lecture: July 1

Examen: July 8

• Makeup: Oct 21 - morning

CONTACT: Lothar Richter richter@rostlab.org

© Burkhard Rost (TUM Munich)
Today: Alignments 1

☐ LAST
  • 3D comparison / Alignment chalk board

☐ TODAY
  • Alignments and “reach of comparative modeling”

☐ NEXT
  • alignment contd
### Notation: protein structure 1D, 2D, 3D

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>kcal/mol</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>40</td>
<td>50</td>
<td>60</td>
<td>70</td>
<td>80</td>
<td>90</td>
</tr>
</tbody>
</table>

**1D**

| P | PP P | 128 | 110 |
| Q | QQQY | 175 | 97  |
| I | FFQVI| 70  | 60  |
| T | SSIVR| 77  | 69  |
| L | LLSTL| 120 | 14  |
| W | WWQED| 238 | 81  |
| Q | RKQAK| 169 | 97  |
| R | RRRFQ| 200 | 62  |
| P | PPPPP | 24  | 48  |
| L | VVT KF | E 71 | E 59 |
| V | VVLII | E 14 | E 0  |
| T | TTKEK | E 74 | E 69 |
| I | AALIV | E 0  | E 0  |
| K | HYK KF| E 90 | E 73 |
| I | IILVI | E 4  | E 0  |
| G | EENGG | 46  | 41  |
| G | GGTTG | 62  | 53  |
| Q | QRRKRR| 68  | 71  |
| L | PPLNW | E 118| E 59 |
| K | TVFKV | E 31 | E 73 |
| E | EESK | E 124| E 95 |
| A | VVG ILG| E 1  | E 0  |
| L | LLI L | E 29 | E 0  |
| L | LLLVV | E 24 | E 0  |
| D | DDDDD | E 49 | E 58 |
| T | TTTTT | 72  | 51  |
| G | GGGGG | 62  | 30  |
| A | AAAAA | 17  | 0   |
| D | DDDDD | 102 | 79  |
| D | DDAKE | 69  | 58  |
| T | SSTTV | 1   | 69  |
| V | IVIVIV | E 14 | E 0  |
| L | VVIVL | E 0  | E 0  |
Sequence comparisons: pairwise methods
Alignments answer: How similar are proteins?
Alignments

OPA
PAPI
Alignments

OPA.. .PAPI
Alignments

OPA..
PAPI
PAPA
Alignments

.PAPI
OPA..
PAPA
.PAPU
.PIPI
Alignments

.PAPI
OPA..
.PAPA
.PAPU
.PIPI

father
Alignments

grandfather

OPA..

.PAPA

.PAPU

.PIPI

father

something else
Correct alignment would have been

grandfather

father

.PAPI
OPA..
.PAPU

.PAPU
PIPE

© Burkhard Rost (TUM Munich)
Alignments: Steps to do

☐ goal: align two sequences

☐ 2 do:
   - (1) Find optimal superposition of the two
   - (2) Define “optimal”
Alignment: Local vs. Global

compare:

- global: all residues aligned
  - GGQLAKEEAL.
  - EGQ..PVEVLP

- PAPI
  - O.PA
Alignment: Local vs. Global

- **global**: all residues aligned
  - GGQLAKEEAL
  - EGQPVEVLP

- **local**: best matches
  - GGQLAKEEALA
  - EGQ..PVEVLP
  - PAPI
  - O.PA
  - OPA..
Alignment: Assumptions about meaning

- Biological meaning of insertion (".")?

GGQLAKEEAL
GQ..PVEVL
Alignment: Assumptions about meaning

- Biological meaning of insertion (".")?
- Historical interpretation (reason to start working on the problem):
  evolutionary relation

GGQLAKEEAL
GQ..PVEVL
Meaning of “homology”?
phylogeny: 
phylo: race/tribe/kind 
genesis: birth 
genos: origin
Phylogenetic Tree Of Life

© Wikipedia
Phylogenetic Tree Of Life

© Wikipedia
Phylogenetic Tree of Life

© Wikipedia
Phylogenetic Tree Of Life

Phylogenetic Tree of Life

Bacteria
- Spirochetes
- Proteobacteria
- Cyanobacteria
- Bacteroides
- Thermotogae
- Aquifex

Archaea
- Methanobacteria
- Methanococcus
- Thermoplasma
- Halophiles
- Entamoeba
- Slime molds
- Animals
- Fungi

Eucaryota
- Plants
- Ciliates
- Flagellates
- Trichomonads
- Microsporidia
- Diplomonads
- Animals
- Slime moulds
- Plants
- Algae
- Protozoa
- Animals
- Fungi
- Gram-positives
- Chloroplastidae
- Green nonsulfur bacteria
- Actinobacteria
- Planctomycetes
- Spirochaetes
- Fusobacteria
- Cyanobacteria
- Chloroplastidae
- Chlorophyll a
- Thermophiles
- Sulfur-reducers
- Acidobacteria
- Proteobacteria
- Crenarchaeota
- Nanoarchaeota
- Euryarchaeota

© Wikipedia
Evolution: speciation

© http://evolution.berkeley.edu/evosite/evo101/
homology
homology definition

- **etymology**: Greek: ομόλογος (homologos): homo (agreeing/same) + logos (word/reason/opinion) (wikipedia: relation)

- **originally** (1843 Richard Owen): “same organ in different animals under every variety of form and function”

- **Genes**: homologs originate from common ancestor
  - orthologs: speciation event
  - paralogs: duplication event
Evolution: speciation

© http://evolution.berkeley.edu/evosite/evo101/

species=?
Evolution: speciation

species=mating

© http://evolution.berkeley.edu/evosite/evo101/
Evolution: speciation

© http://evolution.berkeley.edu/evosite/evo101/

happily munching
Evolution: speciation

© http://evolution.berkeley.edu/evosite/evo101/

disaster strikes
Evolution: speciation

© http://evolution.berkeley.edu/evosite/evo101/

populations diverge
Evolution: speciation

© http://evolution.berkeley.edu/evosite/evo101/

rejoined - yet separated
Evolution: speciation


Evolution: speciation

© http://evolution.berkeley.edu/evosite/evo101/

Happy Face Spider *Theridion grallator*  
(same species-interbreed)

Carrion/Hooded crow  
(same species?)
homology definition

- etymology: Greek: ομόλογος (homologos): homo (agreeing/same) + logos (word/reason/opinion) (wikipedia: relation)

- originally (1843 Richard Owen): “same organ in different animals under every variety of form and function”

- Genes:
  homologs originate from common ancestor
  - orthologs: speciation event
  - paralogs: duplication event

- misuse of term “homolog structure”
  => similar structure
Alignment: Assumptions about meaning

- Biological meaning of insertion (".")?

- Historical interpretation (reason to start working on the problem):
  - evolutionary relation
  - common ancestor
  - homology

GGQLAKEEAL
GQ..PVEVL
Dynamic programming: brute force
How to align sequences?

1. visually

GGQLAKEEAL
GQ...PVEVL
How to align sequences?

☐ 1. visually

A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins

Saul B. Needleman and Christian D. Wunsch
Department of Biochemistry, Northwestern University, and Nuclear Medicine Service, V. A. Research Hospital, Chicago, Ill. 60611, U.S.A.
(Received 21 July 1969)

A computer adaptable method for finding similarities in the amino acid sequences of two proteins has been developed. From these findings it is possible to determine whether significant homology exists between the proteins. This information is used to trace their possible evolutionary development.

The maximum match is a number dependent upon the similarity of the sequences. One of its definitions is the largest number of amino acids of one protein that can be matched with those of a second protein allowing for all possible interruptions in either of the sequences. While the interruptions give rise to a very large number of comparisons, the method efficiently excludes from consideration those comparisons that cannot contribute to the maximum match.

Comparisons are made from the smallest unit of significance, a pair of amino acids, one from each protein. All possible pairs are represented by a two-dimensional array, and all possible comparisons are represented by pathways through the array. For this maximum match only certain of the possible pathways must be evaluated. A numerical value, one in this case, is assigned to every cell in the array representing like amino acids. The maximum match is the largest number that would result from summing the cell values of every pathway.

1. Introduction
Dynamic programming: Global

GGQLAKEEAL
EGQPVEVL...
Dynamic programming: Global

GGQLAKEEAL
EGQPVEVL

GGQLAKEEAL
. EGQPVEVL  .
Dynamic programming: Global

GGQLAKEEAL
EGQPVEVL

GGQLAKEEAL
EGQPVEVL

GGQLAKEEAL
..EGQPVEVL

GGQLAKEEAL
..EGQPVEVL
Dynamic programming concept no ins

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>G</th>
<th>Q</th>
<th>L</th>
<th>A</th>
<th>K</th>
<th>E</th>
<th>E</th>
<th>A</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
allowing for gaps
Dynamic programming concept ins

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>G</th>
<th>Q</th>
<th>L</th>
<th>A</th>
<th>K</th>
<th>E</th>
<th>E</th>
<th>A</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dynamic programming concept ins
Dynamic programming concept ins

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>G</th>
<th>Q</th>
<th>L</th>
<th>A</th>
<th>K</th>
<th>E</th>
<th>E</th>
<th>A</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Dynamic programming concept ins

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>G</th>
<th>Q</th>
<th>L</th>
<th>A</th>
<th>K</th>
<th>E</th>
<th>E</th>
<th>A</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tuesday May 6, 2014
Dynamic programming concept

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>G</th>
<th>Q</th>
<th>L</th>
<th>A</th>
<th>K</th>
<th>E</th>
<th>E</th>
<th>A</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Q</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>P</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

GGQLAKEEAL
GQP..VEVL

GGQLAKEEAL
G.QP.VE.VL
Dynamic programming dot plots

Dot plot of a human zinc-finger transcription factor (GenBank NM_002383) against itself to show self-similarity
what is best?

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>G</th>
<th>Q</th>
<th>L</th>
<th>A</th>
<th>K</th>
<th>E</th>
<th>E</th>
<th>A</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© Burkhard Rost (TUM Munich)
2 of 2 better than 4 of 8?
Gap/Insertions: penalty

- linear gap penalty:

  N gaps cost N

  EGQ.....PVEVLP  <->  E.G.Q.P.V.EVLP
Gap/Insertions: penalty

- **linear gap penalty:**
  
  N gaps cost N

  \[ EGQ\ldots PVEVLP \leftrightarrow E.G.Q.P.V.EVLP \]

- **idea of BLOCKS:**

How to enforce blocks in alignments?
Gap/Insertions: penalty

- **linear gap penalty:**
  
  N gaps cost N

  \[ \text{EGQ} \ldots \text{PVEVLP} \leftrightarrow \text{E.G.Q.P.V.EVLP} \]

- **affine gap penalty:**
  
  gap open
  gap extension (elongation)

  typical go=10 x ge
Dynamic programming: optimal alignment

• Global/no gap:
  SB Needleman and CD Wunsch 1970 J Mol Biol 48, 443-53
• Local/Gap:
  TF Smith and MS Waterman 1981 J Mol Biol 147, 195-197
Dynamic programming: optimal alignment

Pair of protein sequences
U  GGQLAKEEAL
T  EGQPVEVL

Optimal alignment (no gaps)
U  GGQLAKEEAL
T1  EVL
T2  EGQPVEVL

Optimal alignment (with gaps)
U  GGQLAKEEAL
T  EGQP.VE.VL

\[
\begin{array}{ccccccccc}
G & G & Q & L & A & K & E & E & A & L \\
E & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
G & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
Q & 0 & 1 & 2 & 0 & 0 & 0 & 0 & 1 & 1 \\
P & 0 & 0 & 1 & 2 & 0 & 0 & 0 & 0 & 1 \\
V & 0 & 0 & 0 & 1 & 2 & 0 & 0 & 0 & 0 \\
E & 0 & 0 & 0 & 0 & 1 & 2 & 1 & 1 & 0 \\
V & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 & 1 \\
L & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 2 & 1 & 2 \\
\end{array}
\]

- Global/no gap:
  SB Needleman and CD Wunsch 1970 J Mol Biol 48, 443-53
- Local/Gap:
  TF Smith and MS Waterman 1981 J Mol Biol 147, 195-197

\[
SW = \sum_{k=l}^{L_{ali}} M_{U_kT_k} - Go \cdot N_{gap} - Ge \cdot (L_{gap} - N_{gap})
\]
Identity the best criterion?
Substitution matrices
Alignments in brief

>1bl8_A  mol:protein length:97  Potassium Channel Protein
ALHWRAAGAATVLLVLLAGSYLAVLAERGAPGAQLIT
YPRALWWSVETATTGVYGDLYPVTLWGRCVAVVVMVA
GITSFGLVTAAALATWFGREQ

>1orq_C  mol:protein length:223  Potassium Channel
IGDVMEHPLVELGVSYAAALLSVIVVVVECTMQLSGEYLV
RLYLVDLILVILWADYAYRAYKSGDPAGYVKKTLYEI
PALVPGALLALIEGHLAGLGLFRILVRLLRFLRILLHISRG
SKFLSAIAADAADKIRFYHLFGAVMLTVLYGAFAYIVEY
PDPNSSIKSVFDALWWAVVTATTGVYGDVVPATPIGVKVGIAVMLTGISAALTLIGTVSNMFQKILV

Query= 1bl8_A  mol:protein length:97  Potassium Channel Protein
(97 letters)

>1orq_C  mol:protein length:223  Potassium Channel
Length = 223

Score = 58.5 bits (140), Expect = 4e-14
Identities = 26/72 (36%), Positives = 43/72 (59%)

Query: 21  GSYIAYLAERGAPGAQLITYPRALWWSVETATTGVYGDLYPVTLWGRCVAVVVMVAGITS 80
G++ + E P ++ + ALWW+V TATTGVYGD+ P T G+ + + VM+ GI++
Sbjct: 147  GAFAIYIVEYPDPNSSIKSVFDALWWAVVTATTGVYGDVVPATPIGVKVGIAVMLTGISA 206

Query: 81  FGLVTAAALATWF 92
L+ ++ F
Sbjct: 207  LTLLIGTVSNMF 218

© Burkhard Rost (TUM Munich)
Alignments in brief

Query= 1bl8_A  mol:protein length:97  **Potassium Channel Protein**  
(97 letters)

>1orq_C  mol:protein length:223  **Potassium Channel**  
Length = 223

Score = 58.5 bits (140), Expect = 4e-14  
Identities = 26/72 (36%), Positives = 43/72 (59%)

**Scoring matrix**

**Algorithm to optimize score**

**Query: 21**  GSYLAVAERGAPGACLITYPRALWWSVETATTVGYGDLYPVTLLWGRCCAVVVVMVAGITS 80  
G++ + + P + + + + ALWW+V TATTVGYGD+ P T G+ + + VM+ GI++  
**Sbjct: 147**  GAFRIYIVEYPDPNSSIKSVFDALWWAVVTATTVGYGDVVPATPIGKVIGIAVMLTGISA 206

**Query: 81**  FGLVTAALATWF 92  
L+ ++ F  
**Sbjct: 207**  LTLLIGTVSNMF 218
Alignment: Substitution Matrix: PAM

PAM: Point Accepted Mutations (substitution matrix)

Margaret Oakley Dayhoff (1925-1983)

introduced 1978 based on 1572 observed mutations in 71 families of closely related (85% PIDE) proteins

Protein Atlas (Atlas of Protein Sequence and Structure)
(1st sequence: insulin: 1955 Fred Sanger, 1st structures: 1958: Perutz et al. (hemoglobin) / Kendrew et al. (myoglobin) )


PAM1: 1 point mutation per hundred comparisons
PAMn=(PAM1)^n
BLOSUM

- BLOcks of amino acid SUbstition Matrices
  Align only conserved regions
- compile log-odd ratios

\[ S_{i,j} = \log \frac{p_i \cdot M_{i,j}}{p_i \cdot p_j} = \log \frac{M_{i,j}}{p_j} = \log \frac{\text{observed frequency}}{\text{expected frequency}} \]

- BLOSSUMn = threshold at n\% pairwise sequence identity

S Henikoff & Jorja Henikoff (1992)
PNAS 89:10915-9
### Alignments: scoring matrix

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>S</th>
<th>T -2</th>
<th>P -3</th>
<th>A -2</th>
<th>G -3</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>12</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>S</td>
<td></td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td></td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>N -4</th>
<th>D -5</th>
<th>E -5</th>
<th>Q -5</th>
<th>H -3</th>
<th>R -4</th>
<th>K -5</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>M -5</th>
<th>I -2</th>
<th>L -6</th>
<th>V -2</th>
<th>F -4</th>
<th>Y -3</th>
<th>W -8</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>-2</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-3</td>
<td>-4</td>
</tr>
<tr>
<td>I</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
</tr>
<tr>
<td>L</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-4</td>
<td>-6</td>
<td>-7</td>
</tr>
<tr>
<td>V</td>
<td>-2</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-4</td>
</tr>
<tr>
<td>F</td>
<td>-3</td>
<td>-3</td>
<td>-5</td>
<td>-4</td>
<td>-5</td>
<td>-6</td>
<td>-7</td>
</tr>
<tr>
<td>Y</td>
<td>0</td>
<td>-3</td>
<td>-5</td>
<td>-3</td>
<td>-7</td>
<td>-2</td>
<td>-5</td>
</tr>
<tr>
<td>W</td>
<td>-2</td>
<td>-5</td>
<td>-6</td>
<td>-7</td>
<td>-4</td>
<td>-7</td>
<td>-5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>S</th>
<th>T</th>
<th>P</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Alignments: scoring matrix

<table>
<thead>
<tr>
<th>C</th>
<th>S</th>
<th>T</th>
<th>P</th>
<th>A</th>
<th>G</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>Q</th>
<th>H</th>
<th>R</th>
<th>K</th>
<th>M</th>
<th>L</th>
<th>V</th>
<th>F</th>
<th>W</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>N</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>D</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Q</td>
<td>-5</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>H</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>R</td>
<td>-4</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>K</td>
<td>-5</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Scoring Matrix
## Alignments: scoring matrix

### Scoring Matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>H</th>
<th>K</th>
<th>M</th>
<th>N</th>
<th>P</th>
<th>Q</th>
<th>R</th>
<th>S</th>
<th>T</th>
<th>V</th>
<th>W</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
</tr>
<tr>
<td>H</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>K</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>-2</td>
<td>-2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>M</td>
<td>-3</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>N</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>Q</td>
<td>-3</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>R</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>T</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>V</td>
<td>-3</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>W</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>X</td>
<td>-3</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>Y</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
</tr>
<tr>
<td>Z</td>
<td>-3</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
</tr>
</tbody>
</table>

© Burkhard Rost (TUM Munich)
Other substitution matrices

- BLOSUM-62 took the day
- many more (sequence based):
  - GONNET
  - McLachlan
  - Claverie
- structure based
  - PHAT, SLIM
  - STROMA, SDM, HSDM, SM_SAUSAGE, AGAPE
- function based
- for threading/fold recognition
- asf. ....
Interactive software tool

Ignacio Ibarra & Francisco Melo:

Interactive software tool to comprehend the calculation of optimal sequence alignments with dynamic programming

Bioinformatics 2010, 26:1664-5

http:/melolab.org/sat
Dynamic programming

- time used?

- how to choose parameters?
Alignment: hashing (fast & dirty)
BLAST: fast matching of single ‘words’

TTYKLILNLKTLGETTTEAVDAATAEKFQYANDNGVDGEWTYDDATKTFTVTEK
TTYKLILLLLLLLLLLLLLLAWTVEKAFKTFAAAAAAAAWTVEKAFKTFAAAAA

Default “word” size for “seeds” = 3
BLAST: fast matching of single ‘words’

TTYKLILNGKTLKGETTTEAVDAATAEKFVKQYANDNGVDGEWTYDDATKTFVTKEK
TTYKLILLLLLLLLLLLLLLLLLLLAWTVEKAFKTFAAAAAAAAWTVEKAFKTFAAAAA

#1 seed=3

#2 extend

TTYKLIL
TTYKLIL

AATAEKVFQYA
AWTVEKAFKTF

WTYDDATKTF
WTVEKAFKTF

Default “word” size for “seeds” = 3
BLAST: fast matching of single ‘words’

Default “word” size for “seeds” = 3
Minds behind hashing

- **FASTA (extension of FASTN/FASTP)**
  DJ Lipman & WR Pearson (1985)
  Science 227: 1435-41

- **BLAST**
  Basic Local Alignment Search Tool
  SF Altschul, W Gish, W Miller, EW Myers, DJ Lipman (1990) JMB
  215:403-10
BLAST: fast matching of single ‘words’

TTYKLILNGKTLKGETTTEAVDAATAEKFKVFKQYANDNGVDGFWTYDDATKFTVTEK
TTYKLILLLLLLLLLLLLLLLLLLLAWTVEKAFKTFAAAAAAAAATVEKAFKTFAAAAA

#1 seed=3

#2 extend

TTYKLIL
TTYKLIL

AATAEKFVKQYA
AWTVEKAFKTFA

WTYDDATKTF
WTVEKAFKTTF

Default “word” size for “seeds” = 3
the major challenge for word search algorithms is to get the statistics right
Significance of match (e.g. BLAST E-values)
Different approaches for background

- SEARCHSEQ
  - permutations of SEARCHSEQ

FASTA
  - pre-compile distribution for entire database

BLAST

William R Pearson
David Lipman
Stephen Altschul
How accurate are pairwise alignments?
Annotation transfer

similar sequence → similar function

similar sequence → similar structure
Sequence -> Structure

- Sequence folds into unique structure

\[ S \rightarrow T \]

Diagram:
- Structure space
- Sequence space

© Burkhard Rost (TUM Munich)
Sequence -> Structure

- Sequence folds into unique structure: $S \rightarrow T$
- Similar sequences fold into similar structures: $S + S' \rightarrow T$
Sequence -> Structure

☐ Sequence folds into unique structure
  \[ S \rightarrow T \]

☐ Similar sequences fold into similar structures
  \[ S + S' \rightarrow T \]

☐ Most sequences don’t fold, at all
  \[ S \rightarrow \text{no } T \]
Zones

Save Zone

sequences similar

structures similar

0  20  40  60  80  100
Zones

Russell Doolittle
Zones

Midnight Zone

Twilight Zone

Save Zone

sequences similar

structures similar
Schematic: structural similarity in Twilight zone

True Positives = pairs of proteins with similar structure

not true data: just to illustrate the idea of the twilight zone
Schematic: true and false in Twilight zone

True Positives = pairs of proteins with similar structure

False Positives = pairs of proteins with DIFFERENT structure
How to assess alignment accuracy?
How to assess alignment accuracy?
All-vs-all: PDB

1D = sequence alignment
All-vs-all: PDB

3D = structural alignment

1D = sequence alignment
All-vs-all: PDB

3D = structural alignment

1D = sequence alignment

<0.2nm rmsd — SAME 3D
in between — ignore
>0.5nm rmsd — DIFFER in 3D
PDB all-against-all ok?

proteins of known 3D structure (PDB)
Databases biased: MUST remove bias!
Original HSSP curve

HSSP curve revised with 10x data

old Sander & Schneider
new 480*L ^ -0.32 (1 + 1/e^L/1000)

PIDE = Percentage pairwise sequence identity

Number of residues aligned

B Rost 1999 Prot Engin 12, 85-94

© Burkhard Rost (TUM Munich) 96 /128

Tuesday May 6, 2014
Sequence conservation of protein structure

C Sander & R Schneider 1991 *Proteins* 9:56-69
B Rost 1999 *Prot Engin* 12, 85-94
Sequence conservation of protein structure

C Sander & R Schneider 1991 Proteins 9:56-69
B Rost 1999 Prot Engin 12, 85-94
Sequence conservation of protein structure

C Sander & R Schneider 1991 *Proteins* 9:56-69
B Rost 1999 *Prot Engin* 12, 85-94
Structure prediction using homology

C Sander & R Schneider 1991 *Proteins* 9:56-69

Structure prediction using homology

\[ \text{Score} = 83.2 \text{ bits (205), } \text{Expect} = 9e-17 \]
\[ \text{Identities} = 18/101 (X\%), \text{Positives} = 36/101 (35\%), \text{Gaps} = 2/101 (1\%) \]

Query: 111 \text{ AAGGIAAKYLARKNSSVFGFIGCGTQAYFQLEALRRVFDIGEVKAYDREKAKKF 170}
\[ \text{A} + \text{A} + \text{L} + + \text{G} + + \text{L} + + \text{V} + + \text{A} + \]
\[ \text{Sbjct: 153 AAVELAERELGSLHDKTVLVVGAGEMGKTVAKSLVD-RGVRAVLANRTYERAL 211} \]

Query: 171 \text{ EDRGISASVQPAEEASRCDVLVTTPSRKPVVAWVEEGT 211}
\[ + + + \text{R} \text{ D} + \text{V} + \text{ T} + \text{ P} + + \text{ V} + + \text{ E} \]
\[ \text{Sbjct: 212 GGEAVRFDE-LVDHLARSDDVVSATAAPHPVIHVDDVREAL 251} \]

C Sander & R Schneider 1991 *Proteins* 9:56-69
Structure prediction using homology

Score = 83.2 bits (205), Expect = 9e-17
Identities = 18/101 (X%), Positives = 36/101 (35%), Gaps = 2/101 (1%)

Query: 111 AAGGIAAKLARKNSSVFGFIGCGTQAYFQLEALRREKAKKF 170
AA +A +L + +G G + +L + +V + +A +
Sbjct: 153 AAVELAERELGSLHDKTVLVVGAGEMGKTVAKSLVD-RGVRAVLVANRTYERAVEL 211

Query: 171 EDRGISASVQPAEEASCDVLTTPSRKPPVKAEWEEGT 211 + + +R DV+V+ T + PV+ + V + E
Sbjct: 212 GGEAVRFDE-LVDHLARSDVVVSSATAAPHPVIHVDDVREAL 251

C Sander & R Schneider 1991 Proteins 9:56-69
Structure prediction using homology

P1

P2

Score = 83.2 bits (205), Expect = 9e-17
Identities = 18/101 (X%), Positives = 36/101 (35%), Gaps = 2/101 (1%)

Query: 111 AAGGIAAKLARKNSSVFGFIGCGTQAYFQLEALRRVFDIGEVKAYDVREKAACKF 170
  +A + L   +      +G G       ++L     +  V   +   + A +
Sbjct: 153 AAVEELAERELGSLHDKTVLAVGAGEMGKTVAKSLVRVRAVLVANRTYERAVEL 211

Query: 171 EDRGISASVQPAEEASRCDVLVTTPSRKPVVKAEEWEEGT 211
  +       +R DV+V+ T +  PV+ + V E
Sbjct: 212 GGEAVRFDE-LVDHLARSDDLSSATAAHPVIIHVDDVREAL 251

C Sander & R Schneider 1991 *Proteins* 9:56-69

Tuesday May 6, 2014
Structure prediction using homology

Score = 33.9 bits (77), Expect = 0.068
Identities = 14/58 (y%), Positives = 28/58 (48%), Gaps = 2/58 (3%)

Query: 178 SVQPAEEASRCDLVTPTPSRPVKAEBVEEGTHINAGADGPKQELD-VEILKKA 234 + EE ++ D+LV T ++ VK EW++ G ++ G ++ E ++A
Sbjct: 198 TAHLDREVNGDILVVATQPE-MVKGEIKPGAIIVDCGINYKVGVGDAEKERA 254

C Sander & R Schneider 1991 Proteins 9:56-69
Structure prediction using homology

C Sander & R Schneider 1991 *Proteins* 9:56-69

Score = 33.9 bits (77), Expect = 0.068
Identities = 14/58 (24%), Positives = 28/58 (48%), Gaps = 2/58 (3%)
Query: 178 SVQPAEEASRCDELVTTPSRPVKAEWVEETHINAIGADGPQELD-VEILKKA 234
+ EE ++ D+LV T ++ +VK EW++ G + G + ++ E ++A
Sbjct: 198 TAHLDEEVNKGDILVVATGQPE-MVKGEWIKPGAIIVDICGINYKVDVAYDEAKERA 254

Score = 40%, Identities 38 (y%), Positives 28/58 (48%), Gaps = 2/58 (3%)
Sequence conservation of protein structure

C Sander & R Schneider 1991 *Proteins* 9:56-69
B Rost 1999 *Prot Engin* 12, 85-94
How to estimate performance from the curves?
Distance from new HSSP-curve

Sequence identity implies structural similarity!

Distance from curve = +10
Distance from curve = -10

Don't know region

Percentage sequence identity

Number of residues aligned

C Sander & R Schneider 1991 *Proteins* 9:56-69
B Rost 1999 *Prot Engin* 12, 85-94
Distance from new HSSP-curve

C Sander & R Schneider 1991 *Proteins* 9:56-69
B Rost 1999 *Prot Engin* 12, 85-94
Twilight zone = false positives explode!!

Number of protein pairs

Distance from HSSP threshold

Percentage sequence identity

B Rost 1999 Prot Engin 12, 85-94

© Burkhard Rost (TUM Munich)
Twilight zone = false positives explode!!

![Graph showing the relationship between Percentage sequence identity and Distance from HSSP threshold. The graph illustrates a logarithmic scale for both axes, with the y-axis representing the number of protein pairs in.log scale, ranging from 10^0 to 10^6, and the x-axis representing the distance from the HSSP threshold in log scale, ranging from -15 to 10. The graph shows exponential decay as the distance from the threshold increases.]
Detecting true hits in Twilight zone

B Rost 1999 Prot Engin 12, 85-94
Detecting true hits in Twilight zone

Percentage of cumulative true positives

Distance from threshold

they-dont-know-what-they-do only sequence identity

B Rost 1999 Prot Engin 12, 85-94
Finding similar structures in Twilight zone

Number of cumulative true positives

Distance from threshold

acc=9%

acc=15%

acc=59%

B Rost 1999 Prot Engin 12, 85-94
Accuracy vs. Coverage

ACCURACY = \[ \text{how many of the correct proteins were found?} \]

COVERAGE = \[ \text{how many of the proteins found are correct?} \]
BLAST is not enough ...

they don't know what they do
only sequence identity

B Rost 1999 Prot Engin 12, 85-94

© Burkhard Rost (TUM Munich)
Threshold in percentage sequence identity

C Sander & R Schneider 1991 *Proteins* 9:56-69
Threshold in percentage sequence identity

C Sander & R Schneider 1991 *Proteins* 9:56-69
So far: pairwise - anything more 2 do?
Triangle beyond reach

protein A

protein B

protein C
Sequence Space Hopping

protein A

seq_x
seq_y
sel_x

protein B

anb_x
unk_x
unk_y

protein C

cal_x
cal_y
Success through sequence space hopping

![Graph showing percentage sequence identity vs. distance from threshold](image_url)
Sequence comparisons: multiple alignment
Multiple alignments

Dynamic programming?

for 3 sequences: $O(N_1 \times N_2 \times N_3)$

NP-complete (L Wang & T Jiang (1994) JCB 1: 337-48)
Multiple alignments

☐ Dynamic programming?
  for 3 sequences: $O(N_1 \times N_2 \times N_3)$
  NP-complete (L Wang & T Jiang (1994) JCB 1: 337-48)

☐ claim: computer: up to 6
  ~60 TB main memory
  no quote-> unsure

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>G</th>
<th>Q</th>
<th>L</th>
<th>A</th>
<th>K</th>
<th>E</th>
<th>E</th>
<th>A</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Multiple alignments

- Dynamic programming? for 3 sequences: $O(N_1 \times N_2 \times N_3)$
  NP-complete (L Wang & T Jiang (1994) JCB 1: 337-48)

- hack 1:
  dynamic programming: pairwise, only space in vicinity of intersection searched n-wise
Dynamic programming?
for 3 sequences: $O(N_1 \times N_2 \times N_3)$

hack 1:
dynamic programming: pairwise, only space in vicinity of intersection searched n-wise

hack 2:
map to tree / pairwise
Russell Doolittle, UCSD
## Multiple alignment: progressive 1

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>90</td>
<td>80</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>90</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Alignment</th>
<th>Sequence 1</th>
<th>Sequence 2</th>
<th>Sequence 3</th>
<th>Sequence 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGQLAKEEAL</td>
<td>GGQLAKDEAL</td>
<td>GGQIAKDEAL</td>
<td>GGQIAKDEAI</td>
<td></td>
</tr>
</tbody>
</table>

© Burkhard Rost (TUM Munich)
**Multiple alignment: progressive**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>90</td>
<td>80</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>90</td>
<td>80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Step 1**

- GGQLAKEEAL
- GGQLAKDEAL
- GGQLAKDEAI
- ggqlakeeal

---

© Burkhard Rost (TUM Munich)
Multiple alignment: progressive

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>90</td>
<td>80</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>90</td>
<td>80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Step 1**
- GGQLAKEEAL
- GGQLAKDEAL
- ggqlakeeal

**Step 2**
- GGQIAKDEAL
- GGQIAKDEAI
- ggqiaakdeal
Multiple alignment: progressive 1

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>90</td>
<td>80</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>90</td>
<td>80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Step 1:
GGQLAKEEAL
GGQLAKDEAL
ggqlakeeal

Step 2:
GGQIAKDEAL
GGQIAKDEAI
ggqiakdeal

Step 3:
GGQIAKDEAL
GGQIAKDEAI
ggqiakdeal
## Multiple alignment: progressive 2

### Step 1

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>90</td>
<td>80</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>90</td>
<td>80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**GGQLAKEEAL**

**GGQLAKDEAL**

**GGQIAKDEAL**

**GGQIAKDEAI**

---

**Step 1**

- GGQLAKEEAL
- GGQLAKDEAL
- ggqlakeeal
Multiple alignment: progressive 2

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>90</td>
<td>80</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>90</td>
<td>80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Step 1:
- GGQLAKEEAL
- GGQLAKDEAL
- ggqlakeeal

Step 2:
- GGQIAKDEAI
- GGQIAKDEAL

© Burkhard Rost (TUM Munich) 126 /128 Tuesday May 6, 2014
Multiple alignment: progressive 2

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>90</td>
<td>80</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>90</td>
<td>80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Step 1:
- GGQLAKEEAL
- GGQLAKDEAL
- ggqlakeeal

Step 2:
- GGGQLAKDEAI
- GGQIAKDEAL
- ggqlakeeal

Step 3:
- GGGQLAKDEAI
- GGQIAKDEAI
- ggqlakeeal

Tuesday May 6, 2014
Lecture plan PP1: Structure

01: 2014/04/08 Tue: sorry
02: 2014/04/10 Thu: welcome: who we are
03: 2014/04/15 Tue: Intro I - acids/structure (Andrea Schafferhans)
04: 2014/04/17 Thu: SKIP: Easter vacation
05: 2014/04/22 Tue: SKIP: Easter vacation
06: 2014/04/24 Thu: Intro II - 3D comparisons
07: 2014/04/29 Tue: Alignment 1
08: 2014/05/01 Thu: SKIP: “May day” - (NOT to be confused with “m’aidez”)
09: 2014/05/06 Tue: SKIP: student assembly (SVV)
10: 2014/05/08 Thu: Alignment 2
11: 2014/05/13 Tue: Comparative modeling 1
12: 2014/05/15 Thu: Comparative modeling 2
13: 2014/05/20 Tue: Experimental structure determination
14: 2014/05/22 Thu: 3D -> 1D: Secondary structure assignment
15: 2014/05/27 Tue: 1D: Secondary structure prediction 1
16: 2014/05/29 Thu: SKIP: holiday (Ascension Day)
17: 2014/06/03 Tue: Normal mode analysis (Edda Kloppmann)
18: 2014/06/05 Thu: 1D: Secondary structure prediction 2
19: 2014/06/10 Tue: SKIP: Whitsun holidays
20: 2014/06/12 Thu: 1D: Transmembrane helix prediction
21: 2014/06/17 Tue: Nobel prize symposium
22: 2014/06/19 Thu: SKIP: Corpus Christi (Fronleichnam)
23: 2014/06/24 Tue: 1D: Transmembrane strand prediction, solvent accessibility
24: 2014/06/26 Thu: 2D prediction
25: 2014/07/01 Tue: 3D prediction/wrap up
26: 2014/07/03 Thu: examen, no lecture
27: 2014/07/08 Tue: examen, no lecture
28: 2014/07/10 Thu: no lecture